Significance - Royal Statistical Society. 2008, Vol. 5, Issue 1, pp18-24.
(Minor edits in published version)

ANASE: MEASURING AIRCRAFT NOISE ANNOYANCE VERY UNRELIABLY

Peter Brooker

Cranfield University

p.brooker@cranfield.ac.uk

Does anyone who lives under a flight-path like aircraft noise? It is a political hot potato as well: noise annoyance is a factor that a government trying to make the case for a third runway at Heathrow must deal with – tens of thousands of people will hear its noise. So, when a study commissioned by the Department for Transport claimed that people are becoming less tolerant of aircraft noise than they used to be, it was, to say the least, highly unpleasant reading for supporters of a third runway. But the DfT had major concerns about the report. Peter Brooker senses the vibrations.

1. Aircraft: Too Loud and Too Many

We almost all like to fly, but few of us want to live under an airport flight-path. The noise that planes inflict affect the quality of life, to say nothing of house-prices, in the neighbourhoods beneath. Proposals for changes to airports, such as the third runway for Heathrow, have to include descriptions of how they will change noise exposure – although estimating the total noise from aircraft over the course of a day is actually very complex. In the UK, noise contours around airports use an index called Leq, short for Equivalent Continuous Sound Level, which is essentially the noise energy received on the ground. It is measured in decibels; it takes into account both the noise levels of aircraft and their number, and averages the sound energy from all aircraft movements in a certain area over a 16 hour period each day, between seven in the morning and eleven at night. (Sleep disturbance from night flights is a separate UK policy concern – but an equally, if not more, serious issue for many people.)

The UK Department for Transport [DfT] has used Leq following the 1985 Aircraft Noise Index Study (ANIS)[1,2] conclusion that there was no better metric than Leq in terms of correlation between aircraft noise and community annoyance. The Government decided to adopt the use of Leq to describe noise, and decided that 57 Leq marks the approximate onset of significant community annoyance from aircraft noise. 57 Leq has often been termed Low Community Annoyance, Medium Annoyance is 63 Leq, and 69 Leq registers on the scale as High Annoyance.

In mid-2001, the DfT announced a major study into aircraft noise: They subsequently commissioned commercial contractors (led by MVA Consultancy Ltd) to conduct the ANASE (Attitudes to Noise from Aviation Sources in England) project to investigate, among other things, the relationship between aircraft noise and annoyance

> '…the new study underlines the Government's commitment to underpin our policy on aircraft noise by substantial research that commands the widest possible confidence' they said. .'

They hoped this new research would command such confidence. They added that conclusions from the original ANIS research had been

> '…broadly confirmed by other studies here and abroad, and we have no reason to doubt their validity.

Six years later, in November 2007, the ANASE report was published. It concluded that our tolerance of aircraft noise was decreasing; that similar volumes of noise annoy us more now than they used to. For a government considering, for example, the need for a third runway at Heathrow, this was not good news.

The report's 'quantitative findings were rejected as unreliable' [BBC] by the DfT. But about a quarter – *sic* – of the project's duration had in fact been spent on Expert Peer Reviews. ANASE's website (http://www.dft.gov.uk/pgr/aviation/environmentalissues/Anase/) includes the report, its technical appendices and several of these critical reviews.

In particular, DfT paid two acoustics experts to review the ANASE draft material[3]. Their comments include:

> "…in the first version of this review it was stated that there were sufficient technical and methodological uncertainties still remaining with the study to mean that reliance on the detailed outcome of ANASE would be misplaced. In view of developments since the review of the July 2007 version of the ANASE main report, the reviewers are even more convinced that their concerns are fully justified…"

So was the Government right to be concerned about the ANASE claims? What were the actual problems found by these Peer Reviewers? We can summarise the main ANASE claims, and then examine the design, methodology and statistical analyses that its authors set out in their published material.

2. ANASE's Claims

ANASE adopted several basic ideas from its predecessor, ANIS. It used social survey questionnaires to elicit respondents' annoyance from aircraft noise. It studied fifty-six survey sites near nine airports, with levels of noise from 36 to 68 Leq. The study report makes a number of aircraft annoyance claims, including these (slightly edited):

> *Claim*: "For the same amount of aircraft noise, measured in Leq, people are more annoyed in 2005 than they were in 1982."
>
> *Claim:* "The modelling work also showed that respondents were less sensitive to changes in sound level below 42 Leq and above 59 Leq, adding support to a logistic dose-response form [*dose = noise measure, response = annoyance*]. There was no threshold, or discontinuity, in the relationship between mean annoyance and Leq."

A third claim was that Leq itself was a faulty guide to measuring noise:

Claim: "The results … suggest that Leq gives insufficient weight to aircraft numbers, and a relative weight of 20 appears more supportable from the evidence than a weight of 10, as implied by the Leq formulation."

These are dramatic claims. To meet the DfT criterion that they should 'command the widest possible confidence', they would need to be robust, technically reliable, and capable of withstanding scrutiny. But were the ANASE claims actually strong enough to do that?

3. ANASE Problems: Questionnaire

When carrying out an attitudinal survey, you have to make choices about how you word the questions, the context in which you ask them, the scales you assign to the answers you get, and your data collection technique. But all these choices can generate errors and biases. The responses to attitudinal questions may easily be affected by the way the issue is posed, the order of the questions, the particular wording of a key question and its context.

Psychologists interpret attitudes as 'structures in long-term memory'. They suggest that when we are asked about our attitudes, we go through a four-stage cognitive process before we arrive at an answer:

(i) Interpret the question ("What is the attitude about?").

(ii) Retrieve relevant beliefs/feelings.

(iii) Apply these beliefs/feelings to generate appropriate judgement.

(iv) Use this judgement to formulate response.

This indicates that attitudes are not some kind of enduring personal view, just waiting to be picked out of someone's mind. Instead, they are 'evaluative judgements' formed at a particular time – and formed both by the questions that are being asked, and the way in which they are being asked.

Thus, attitude reports – how people 'feel' about certain subjects – are highly context sensitive. All four stages above can potentially be affected by 'prior items': serious respondents may be building on their earlier thought processes, or they may aim to 'match' the earlier responses and be consistent with their answers. They are unlikely to want to mislead about their 'true' attitudes, but they may be motivated to provide answers showing that they are aware of the issues about which they are to be questioned.

The message is clear: to make valid comparisons of attitudes over time, you have to ask exactly the same questions in the same at each time point. This is standard textbook guidance[4] repeated by (eg) the UK Government Social Research Unit [GSRU][5] and in NHS research guidelines[6]. So, if ANASE wanted to find how our attitudes to noise had changed since 1982, it should have asked the same questions, in the same context, as were asked in 1982. It did not.

Figure 1 shows a schematic comparison of the ANIS and ANASE questionnaire set-ups. Two potential context effects are:

ANASE installed noise playback equipment in people's homes before interviewing them; ANIS did not. Thus, ANIS is a social survey and ANASE is

a combination of a social survey and a laboratory experiment – later in the interview, noises are played to respondents.

ANASE starts immediately with questions on aircraft noise annoyance; but ANIS leads up to them by asking first about perceptions of the local area, hence allowing the interviewee to mention aircraft noise spontaneously.

Both of these factors could affect annoyance ratings considerably. And other factors can have a major impact too – such as recent media attention on an airport's possible growth, and people's trust in both the airport company and in national/local airport government policies. .

5. ANASE Problems: Annoyance Measure

The ANASE contractors' way of using annoyance scales is odd. First, compare the questions that ask a respondent how much he or she is annoyed:

ANIS	**ANASE**	
Very much?	Extremely?	Highly Annoyed?
	Very?	
Moderately?	Moderately?	
A little?	Slightly?	
Not at all?	Not at all?	

Note that the ANIS version has no middle ranking choice – so the interviewee is not able to take the 'easy way out' by choosing 'in the middle'. For the ANASE version, the combination of 'Very' and 'Extremely' answers is taken as a 'Highly Annoyed' category. The ANASE reports did not offer evidence-based reasons for the change.

There is no perfect recipe for determining 'good' attitude scales, but the key question is the extent to which a possible scale is cardinal in nature (ie corresponding to the properties of integers), rather than just being 'ordinal' (ranking responses). If a scale is cardinal, then such results can be manipulated by all the rules of arithmetic, and hence analysed by the standard kinds of statistical testing.

ANIS used the responses above to construct a 'Very Much Annoyed Percentage' scale of annoyance at each survey site. The ANIS choice of scale is consistent with the great bulk of world-wide research on aircraft disturbance (eg Fidell & Silvati[7], a recent international review paper of social survey data into aircraft noise annoyance). In contrast, ANASE used the answers to its version of the annoyance question to construct a 'Mean Annoyance'. In its scheme, a rating of 'Not at all' scored 10 points, of 'Slightly' scored 30 points', up to 'Extremely' scoring 90 points; ie each extra level of annoyance added twenty points. The Mean Annoyance estimate for the site was then simply the arithmetic average of the respondents' scores, eg if half the people said 'Not at all' and half the people said 'Extremely', this would be a mean of 50 points.

But ANASE's choices of weightings are subjective value judgements. The ANASE contractors did not produce robust evidence to justify the relative numerical scorings

(saying the scale is ‘standardised’ adds no content). Why are nine people saying ‘Not at all’ equivalent to one person saying ‘Extremely’, or to three people saying ‘Slightly’? Rather than 10, 30, 50, 70, and 90, the analysis could have used any other set of increasing numbers – and that would have changed the inferences they made.

Arbitrarily averaged attitude scales, with their unreliable statistical properties, were used very cautiously even before the ANIS work. It is puzzling why ANASE would need to change from the ANIS percentage scales, of how many were Very much annoyed, how many Moderately, and so on. But even when we analyse ANASE in percentage terms, so that it is comparable with ANIS and international work, problems remain.

6. ANASE Problems: Statistical Analysis

ANASE used two kinds of survey sites. At one (‘Full’) there was the noise playback equipment of Figure 1, and at the other (‘Restricted’) there was no equipment. Thus, the context for the two was markedly different. If context effects are crucial in this study, then marked differences would be expected in the data from the two kinds of sites – and they are there.

Figure 2 shows the ‘% Highly Annoyed’ response for the two site types at the 27 ANASE Heathrow sites. The Heathrow sites are selected because of the availability of CAA [Civil Aviation Authority] / DfT higher accuracy Leq values for these sites; because it is straightforward to approximate internationally-used DNL values (by adding 2.5 to the Leq value); and to avoid airport-dependent factors. [DNL is the Day-Night Average Sound Level used in the USA and several other countries: it is a 24-hour Leq with night flight noise levels artificially increased by 10 decibels.] Simple linear-fit trend lines are also shown for the two sets of data.

Figure 2 indicates that the Full and Restricted scatter plots and trends are very probably different – in particular the trend line slopes differ. In comparing two regression lines, the most basic hypothesis to test is the hypothesis of coincidence, ie if the two underlying relationships are the same. The ANASE contractors carried out statistical testing to compare Heathrow Full and Restricted data – but only at the instigation of the reviewers[3]. This rejected the coincidence hypothesis, finding that the differences were statistically significant (t-statistic above the standard 5% level). It is therefore unlikely that the two samples come from the same underlying population. It implies that the introduction of noise equipment changed the aircraft noise annoyance dose-response relationship, by a roughly multiplicative bias. The ANASE contractors decided to ignore these crucial results.

Only in circumstances when statistical testing accepts coincidence, as examined through (eg) Analysis of Variance techniques, is it permissible to fit a single overall regression line to both relationships. But the ANASE statistical analysis wrongly combines Full and Restricted data sets (eg Figure 3). To ignore the statistical testing results rejecting the coincidence of the data sets is not sound practice. It is the kind of thing that a statistical textbook would offer as an example of ‘how to do it incorrectly’. It removes any sound foundations for subsequent ANASE modelling claims about (eg) annoyance onsets and the weighting of the number of aircraft.

Why do the Full and Restricted data sets differ? It is not possible to offer precise reasons based on the ANASE documents, simply because the ANASE work did not investigate potential causes. One factor could be confusion between hearing and/or being aware of noise as compared with suffering a degree of annoyance. The presence, and presumed intended use of the noise playback equipment, is certainly a possible strong factor.

An even more telling illustration is a mapping of the Heathrow data in Figure 2 onto the Fidell & Silvati[7] data set – Figure 4. This aircraft annoyance research collated international data from 326 site surveys with an average of about 160 people per site. The Figure shows a scatter plot of all the '% Highly Annoyed' data against DNL. The two trend lines are the linear fits to the Fidell & Silvati data and the ANASE Heathrow Full data. The ANASE Heathrow Restricted data lies roughly on the Fidell & Silvati trend line. The ANASE Heathrow Full data lies markedly above the trend line for the other data: it is hard to believe that it is a sample from the same underlying population.

Figure 5 shows the complete set of Full and Restricted data from ANASE (using wholly ANASE data). This again shows that there are differences between the two data sets: having noise equipment present does make a difference – showing a roughly multiplicative bias at the Full sites. The Figure also shows that ANASE Restricted sites were not wisely selected. The onus was on the ANASE contractors to select sites to be able to test effectively for Full/Restricted differences – Restricted sites at higher Leq values ('control group sites') should therefore have been included.

Figure 6 compares the '% Highly Annoyed data' from all the Restricted sites with a curve fitted to the ANIS results used in policy work[3] (Fidell and Silvati[7] discuss curve-fitting). The ANASE Restricted data points are possibly slightly above the ANIS curve, but this could be a statistical sampling issue (Restricted site ANASE samples were very small, typically 16 people) and/or a context effects-related problem – because of a markedly different questionnaire ordering and a different annoyance question.

7. ANASE Problems: International Comparisons over Time

There are comments in the ANASE reports that allude to non-UK studies suggesting that the annoyance dose-response relationship might be moving upwards, ie people are typically more annoyed for a given Leq. This is not a new suggestion[1]. The test of this kind of hypothesis is to examine data.

As already noted, an excellent recent review paper is Fidell & Silvati[7]. Figure 7 extracts results from the Fidell & Silvati data set. It shows responses in the bands 47.5-52.5, 52.5-57.5, and 57.5-62.5; ie these represent ~50, ~55 and ~60 DNL. The plots cover results after 1980, mainly because the interest is in changes since the early 1980s ANIS work. The Figure plots these responses against the year the survey was published. Simple (unweighted) linear regressions on the data in the Figure – the trend lines – do not show significant changes over time (none of the regression t-statistics is significant at even the 10% level). Thus, there is no strong evidence from this large international data set of any trend over time.

A simple analysis on even this large data set is not statistical proof. To be confident about the magnitude of possible trends over time, it would be necessary to carry out high-quality data collections and statistical analyses, with tight experimental controls on questionnaire context/design, annoyance scales, socio-economic variables, media attention/trust, and sampling variations.

8. Summary

The DfT was wise to commission the peer reviews and to publish the material rather than be accused of a 'cover up'. But no reliance can be put on ANASE claims: they cannot 'command the widest possible confidence'. There are unrepairable major problems with questionnaire design and process, analysis techniques, and selective attempts to compare with international work.

The design of the ANASE questionnaire does not meet the necessary criteria set out in standard textbooks, by the Treasury's GSRU, or responsible UK organisations (eg the NHS). This damages the ability to make reliable comparisons with earlier work.

The analysis techniques used in ANASE do not recognise the problems of using average annoyance scales in parametric statistical analyses. ANASE's contractors presented no good reasons for changing from earlier, robust scales, *inter alia* preventing proper comparisons.

ANASE fails to meet minimum data analysis requirements for such a study, ie critical examination of raw data to detect potential biases, and always taking proper account of statistical testing results. The regression-based statistical modelling used in ANASE is invalid because it too quickly combines data from Full and Restricted (ie without noise playback equipment) sites samples. This also reveals ANASE's poor design: the onus was on the contractors to test key hypotheses on these effects – there are insufficient Restricted 'control group' sites.

ANASE data suggest that the introduction of noise equipment changes the aircraft noise annoyance dose-response relationship by a roughly multiplicative bias factor – but no points are awarded for measuring the wrong thing, accurately or not. ANASE data for Full sites are markedly out of line with the results of reputable international and previous UK work. As data from ANASE's Full sites are unlikely to be representative of people's annoyance attitudes, the SP results that build from these distorted attitudes may similarly be distorted. ANASE Restricted site data are broadly consistent with international and ANIS results.

Thus, a straightforward factual explanation for the ANASE data set is that it has a design-induced multiplicative bias overlaying annoyance responses largely unchanged from past studies. The implication is that the ANASE contractors' claims – eg increased annoyance over time, additional aircraft number effects – are invalid because they mostly derive from the biased data.

References

1 Brooker, P. (2004). The UK Aircraft Noise Index Study [ANIS]: 20 Years On. Acoustics Bulletin. May/June, 10-16. https://dspace.lib.cranfield.ac.uk/handle/1826/1004

2 Brooker, P., Critchley, J. B., Monkman, D. J. & Richmond, C. (1985). United Kingdom Aircraft Noise Index Study (ANIS): Main Report DR Report 8402, for CAA on behalf of the Department of Transport, CAA, London.

3 Havelock, P. & Turner, S. W. (2007). Attitudes to Noise from Aviation Sources in England: Non SP Peer Review. Environmental Research & Consultancy, CAA; Bureau Veritas. http://www.dft.gov.uk/pgr/aviation/environmentalissues/Anase/nonsppeerreview.pdf

4 Sudman, S. & Bradburn N. M, (1982). Asking Questions: A Practical Guide to Questionnaire Design. San Francisco, Jossey-Bass.

5 Government Social Research Unit (2007). The Magenta Book: Guidance Notes for Policy Evaluation and Analysis. HM Treasury, UK. http://www.policyhub.gov.uk/magenta_book/

6 McColl, E., Jacoby, A., Thomas, L., Soutter, J., Bamford, C., Steen, N., et al. (2001). Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Health Technology Assessment [HTA] 5(31). [NHS R&D HTA Programme]. http://www.hta.ac.uk/fullmono/mon531.pdf

7 Fidell, S. & Silvati, L. (2004). Parsimonious alternative to regression analysis for characterizing prevalence rates of aircraft noise annoyance. Noise Control Engineering Journal, 5(2), March/April, 56-68.

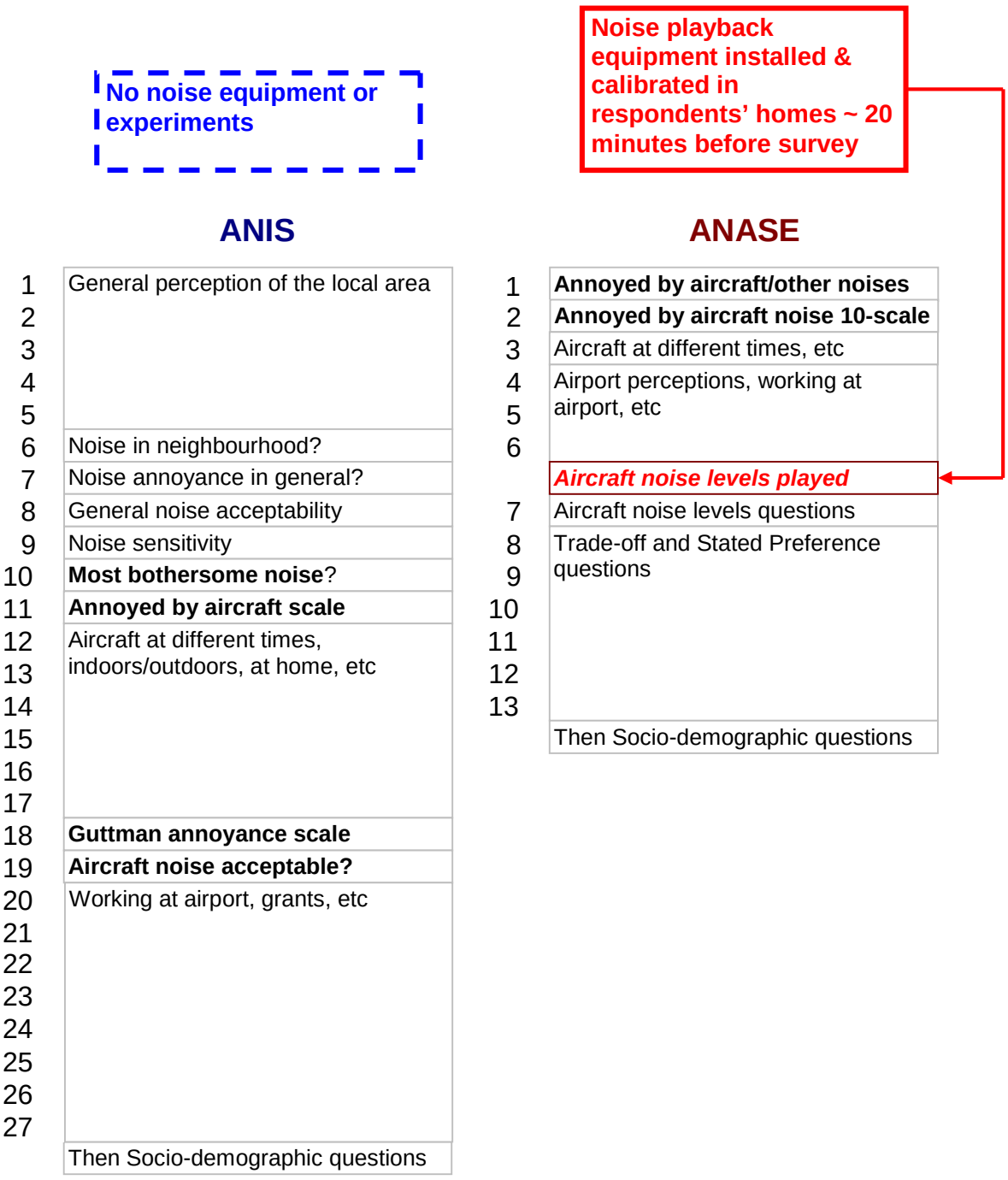

Figure 1. Comparison of key ANIS and ANASE Questionnaire context, question order and noise playback equipment differences

Notes:

(i) The ANASE questions are in the order given, but the number starts at 6 rather than 1 – no explanation is given for this.

(ii) The bold text indicates where questions to provide 'aircraft disturbance' scales used in the statistical analyses were asked.

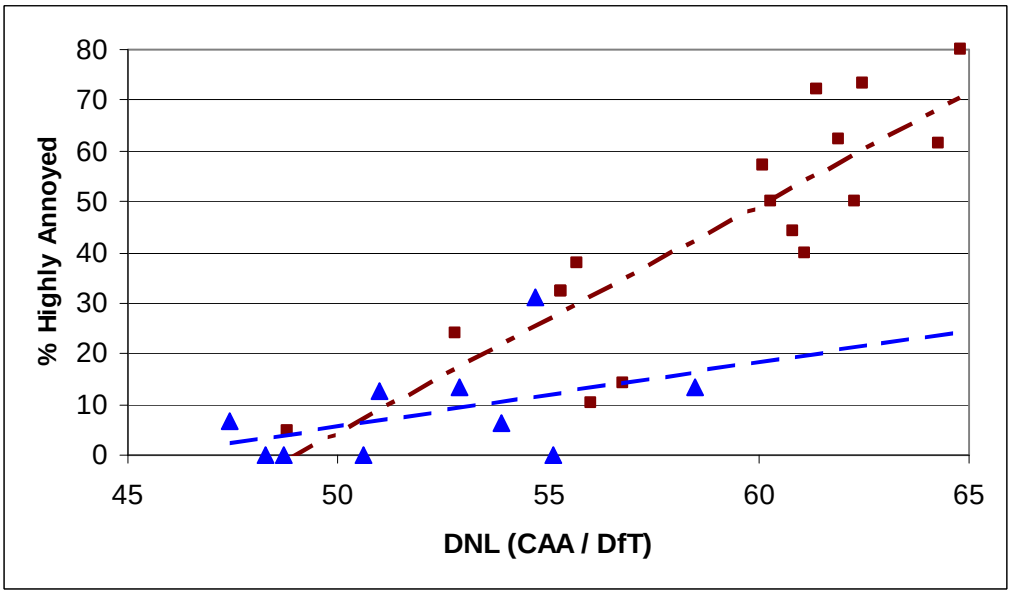

Figure 2. ANASE '% Highly Annoyed' Heathrow results are from two distinct data sets.
Red square – Full; Blue triangle – Restricted. Linear trend lines.

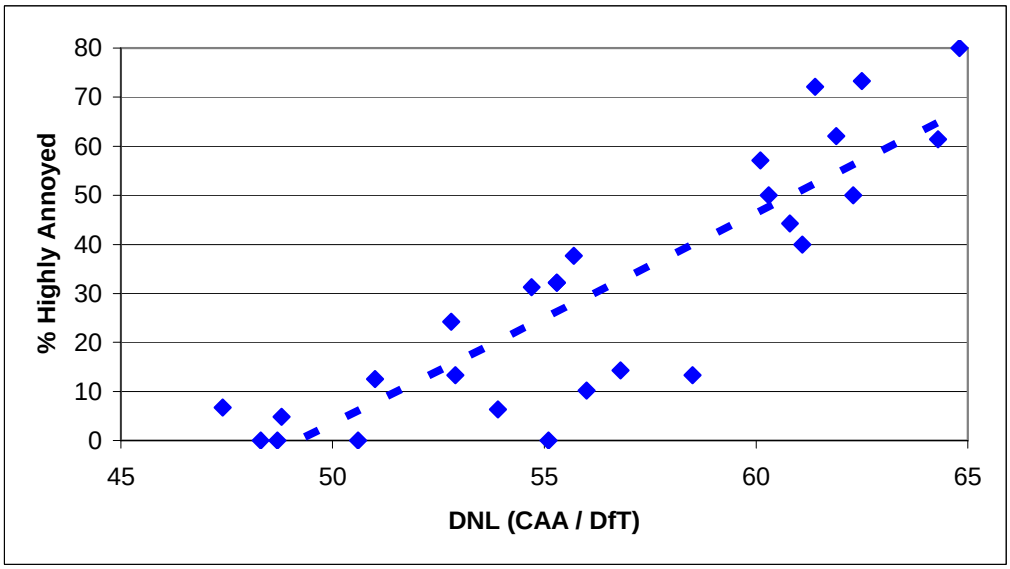

Figure 3. Erroneous ANASE-type fit for Heathrow results – statistical test results disregarded.

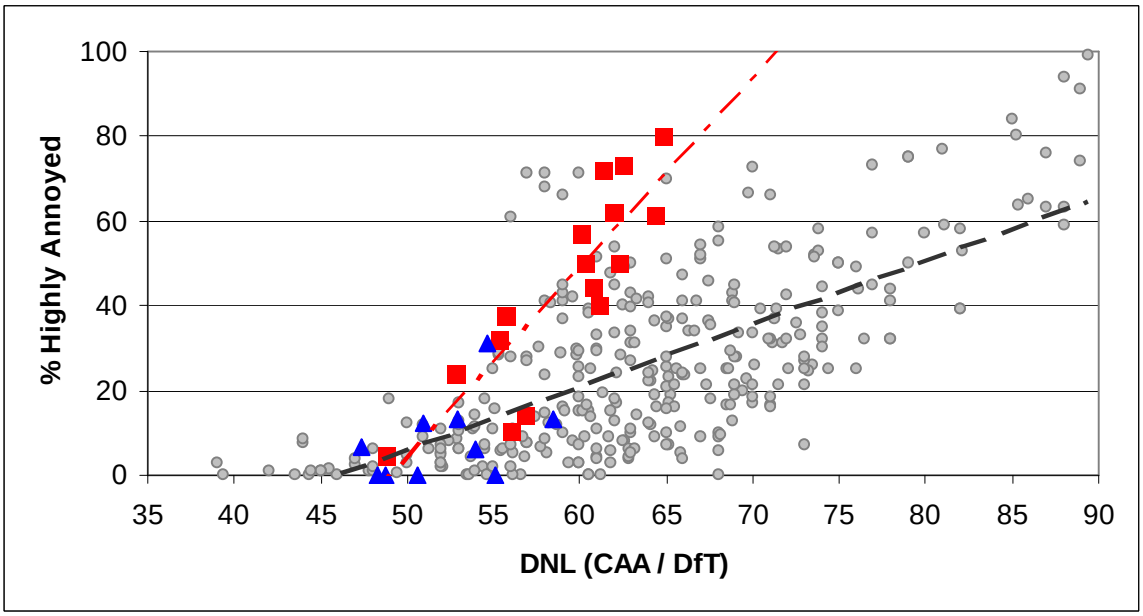

Figure 4. Compares Heathrow ANASE '% Highly Annoyed' with Fidell & Silvati[7]
Red squares – Heathrow Full; Blue triangles – Heathrow Restricted; Grey blobs – Fidell & Silvati data set. Linear trend lines to Full and Fidell & Silvati data.

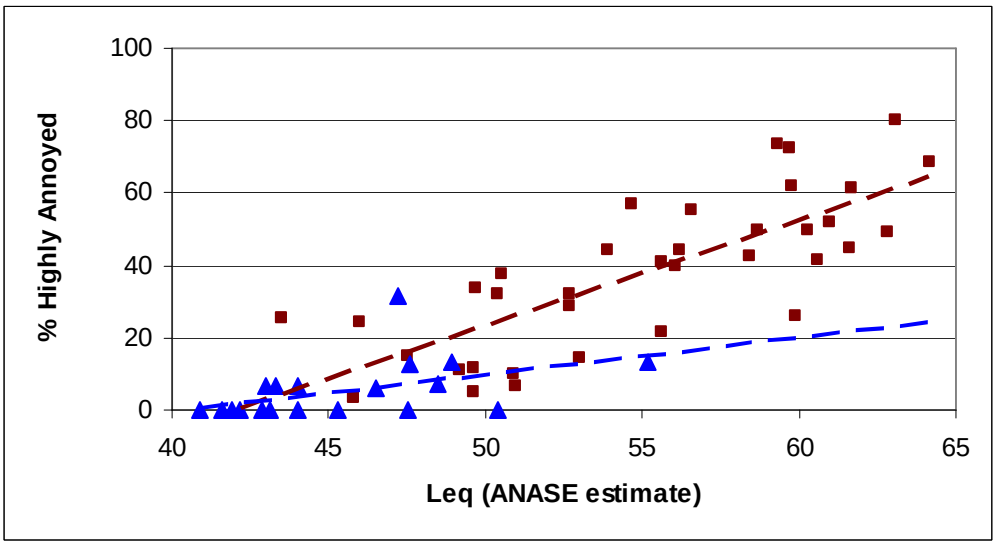

Figure 5. Compares ANASE Full and Restricted sites '% Highly Annoyed'.
Red squares – ANASE Full sites; Blue triangles – ANASE Restricted sites. Linear trend lines. Source Technical Appendices, Table 10 (pages 250/1), Table 6.2 (pages 17/18). Site R17 excluded – as in ANASE analyses.

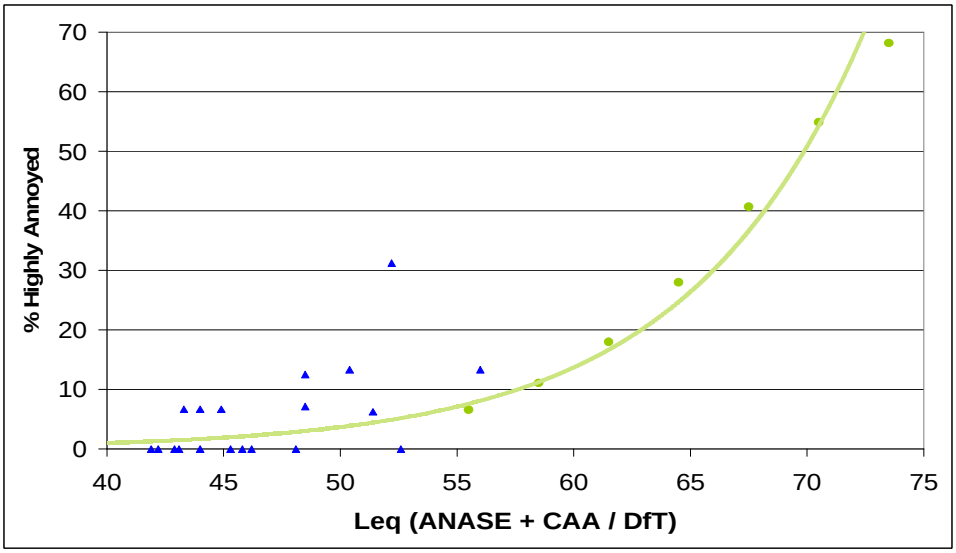

Figure 6. '% Highly Annoyed' at ANASE Restricted sites compared with ANIS curve
Blue triangles – ANASE Restricted sites (source above), sample size typically 16.
X-axis ANASE Leq for non-Heathrow data and CAA / DfT Leq for Heathrow data
Blobs are standard ANIS values[3] (Table 2), plus exponential fit.

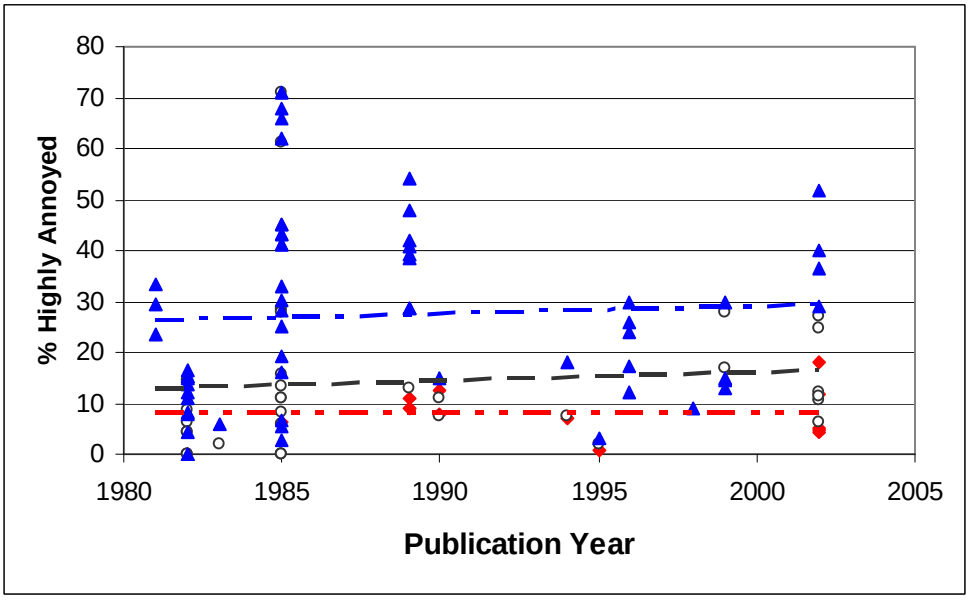

Figure 7. '% Highly Annoyed' from Fidell & Silvati[7], post 1980 data
Red lozenge - ~50 DNL; Open circles - ~55 DNL; Blue triangles - ~60 DNL. Linear trend lines.